W432–W437 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki441

P-Match: transcription factor binding site search by combining patterns and weight matrices

D. S. Chekmenev, C. Haid and A. E. Kel*

BIOBASE GmbH, Halchtersche Strasse 33, D-38304 Wolfenbüttel, Germany

Received February 14, 2005; Revised and Accepted March 30, 2005

ABSTRACT

P-Match is a new tool for identifying transcription factor (TF) binding sites in DNA sequences. It combines pattern matching and weight matrix approaches thus providing higher accuracy of recognition than each of the methods alone. P-Match is closely interconnected with the TRANSFAC® database. In particular, P-Match uses the matrix library as well as sets of aligned known TF-binding sites collected in TRANSFAC® and therefore provides the possibility to search for a large variety of different TF binding sites. Using results of extensive tests of recognition accuracy, we selected three sets of optimized cut-off values that minimize either false negatives or false positives, or the sum of both errors. Comparison with the weight matrix approaches such as Match™ tool shows that P-Match generally provides superior recognition accuracy in the area of low false negative errors (high sensitivity). As familiar to the user of Match™, P-Match also allows to save user-specific profiles that include selected subsets of matrices with corresponding TF-binding sites or user-defined cut-off values. Furthermore, a number of tissue-specific profiles are provided that were compiled by the TRANSFAC® team. A public version of the P-Match tool is available at http://www.gene-regulation.com/cgi-bin/pub/programs/pmatch/bin/p-match.cgi.

INTRODUCTION

Understanding mechanisms of regulation of gene expression on the level of transcription is one of the key problems in post-genomic era. Each cell type or tissue, at a specific developmental stage, in a specific cell cycle phase and under influence of extracellular signals is characterized by a particular pattern of activated transcription factors (TFs). Binding of these nuclear proteins to their cognate binding sites in the regulatory regions (e.g. promoters, enhancers) of genes and subsequent recruitment of co-factors and components of nucleosome remodelling machinery enables formation of multi-protein/DNA complexes that provide gene activation or repression. Therefore, computational methods of predicting TF binding sites in DNA are very important for understanding the molecular mechanisms of gene regulation. Over the past few years, numerous tools have become available for the prediction of TF binding sites [for recent reviews, see (1,2)]. Especially popular are those tools which use information on known TF binding sites that are collected in databases such as TRANSFAC® (3). Approaches vary between high generalization provided by weight matrices to high specialization provided by pattern matching approaches. More sophisticated approaches include consideration of nucleotide correlation in different positions of the sites, HMMs, taking into account flanking regions and some others (4–13). But usually, complex approaches require large training sets, which is rather problematic for the most known TFs for which only small sets of binding sites are known (up to 10 sites).

We have developed a novel tool called P-Match for searching putative TF binding sites in DNA sequences. It effectively combines pattern matching and weight matrix approaches. P-Match uses the TRANSFAC® library of weight matrices as well as sets of aligned known TF-binding sites collected in the TRANSFAC® database. The P-Match algorithm searches for DNA subsequences matching one of the known TF binding sites from each set. It calculates the matching score using the corresponding weight matrix, which is the main novelty of the algorithm in contrast to the classical pattern matching approaches that compute a score by counting just the number of mismatches.

We performed extensive testing of P-Match on sets of genomic TF binding sites, on chromosome sequences and on artificially simulated random sequences in order to compare the recognition accuracy of this method versus classical weight matrix methods and classical pattern matching techniques. Comparison with the weight matrix approaches such as Match™ tool shows that P-Match generally provides superior

*To whom correspondence should be addressed. Tel: +49 5331 8584 41; Fax: +49 5331 8584 70; Email: ake@biobase.de
Present address:
C. Haid, Universität des Saarlandes, Saarbrücken, Germany

recognition accuracy in the region of low false negative errors (high sensitivity). A public version of the P-Match tool is available at http://www.gene-regulation.com/cgi-bin/pub/programs/pmatch/bin/p-match.cgi.

Results of testing are given in the Supplementary Material.

ALGORITHM

The algorithm is based on simultaneous use of a positional weight matrix (PWM) and a set of aligned TF binding sites used to construct this matrix. To construct a matrix in TRANSFAC, we first compile a set of sites by grouping all sites from the database which match a chosen 'experimental evidence quality criteria' and which are known targets of a selected TF or a family of similar TFs (sites for orthologous factors of mammalian species are grouped together). On the next step, we align sites by using a combination of Gibbs site sampling method (14) and a recursive application of Match program (15) ensuring that the 'core' of each site, which is often supported by experimental evidences, is included in the alignment. At last, we choose such a window (i, j) for matrix construction ($1 \leqslant i, j \leqslant$ length of the alignment; and $w = j - i > 6$ bp), which provides the lowest false positive (FP) level for a selected sensitivity level (often chosen as 50%). These tests are done with a leave-one-out procedure, which helps to smoothen the small sample effects.

The PWMs are computed from the nucleotide frequency matrix using the following formula:

$$w_{i,B} = f_{i,B} \times I_i; B \in \{\mathrm{A,C,G,T}\},$$

where

$$I_i = \sum_{B\in\{\mathrm{A,C,G,T}\}} f_{i,B} \log_2 f_{i,B} - \sum_{k=1,4} \frac{1}{4} \log_2 \frac{1}{4},$$

where, $f_{i,B}$ is the frequency of observing nucleotide B in the position i of the alignment, and $w_{i,B}$ is the corresponding element of the weight matrix. I_i is the maximal amount of information in the position i of the matrix which is calculated as a difference between entropy of this position and the entropy of evenly distributed four nucleotides.

The P-Match search algorithm utilizes PWMs from TRANSFAC and computes d-score value which measures similarity between a sub-sequence X of the length L in DNA and a given TF site S from the site set Ω ($S \in \Omega$). The d-score is calculated using weights of the nucleotides in the individual positions of the site taken from the corresponding weight matrix:

$$d = \frac{\mathrm{MaxWeight} - \sum_{i=1,L} |\, w_{i,B(X_i)} - w_{i,B(S_i)} \,|}{\mathrm{MaxWeight}},$$

$$\mathrm{MaxWeight} = \sum_{i=1,L} \max_{B\in\{\mathrm{A,C,G,T}\}} (w_{i,B}),$$

where $B(X_i)$ and $B(S_i)$ are the nucleotides in ith position of the subsequence X and the site S, respectively. The d-score ranges from 0.0 to 1.0, where 1.0 denotes an exact match of the nucleotide weights of the site S to the corresponding weights of the sub-sequence X. Similar to the Match algorithm, we compute two separate d-scores: d_{matrix}—for the whole site and d_{core}—for the core positions of the site, which are the five most conserved positions in the alignment. In the search, we choose two independent cut-off values for these two d-scores (in each site set, the cut-offs are the same for all sites in the set). Only those matches are reported whose both d-scores d_{matrix} and d_{core} exceed the corresponding cut-offs.

The sets of aligned sites contain many sites that are similar to each other. Therefore, the straightforward application of the P-Match algorithm to a DNA sequence may produce several matching of different sites from the set to the same place on DNA. Such redundancy can be removed from the output by invoking a special (default) option of the program, which outputs then one match with the highest d-score out of several overlapping matches. This is very useful for reducing the output, although for some application the information on such multiple matching can help to reveal the most promising potential TF binding sites. In addition, the high similarity of a match in the sequence under study to an existing known site in a promoter of another gene can give an idea about the function of the found potential binding site.

Estimation of optimal *d*-score cut-offs

Selection of the cut-off values largely depends on the user's objectives. Similar to (15), we have pre-calculated three different cut-offs for each TRANSFAC matrix that has the site alignment set attached to it: (i) to minimize false negative rate (under-prediction error); (ii) to minimize false positive rate (over-prediction error), (iii) to minimize the sum of both errors.

Cut-offs minimizing false negative rate (minFN). We use leave-one-out method for estimation of the false negative rate of site recognition. Iteratively, from a site set Ω, we remove one site $S^{(r)}$ ($1 \leqslant r \leqslant |\Omega|$). The reduced set $\Omega\backslash\{S^{(r)}\}$ is used then to run P-Match in order to recognize the removed site $S^{(r)}$. On the basis of these runs, we set the minFN cut-offs to such values that provide recognition of at least 90% of the removed sites $S^{(r)}$. We decided to tolerate an error rate of 10%, taking into account that the set of sites might contain 'weak' representatives.

Cut-offs minimizing false positive rate (minFP). We have applied the algorithm described above to the sequences of the third exons of human genes ($\sim 2 \times 10^6$ bp) because these sequences are presumed to contain no biologically relevant TF binding sites. For every matrix, the lowest cut-off for which no match is found in the set of exon sequences is set to be the minFP cut-off. Since the selection of the background sequences can influence the cut-off selection, we are going to evaluate the use of other genomic sequences to make alternative minFP cut-off estimates.

Cut-offs minimizing the sum of both errors (minSum). We compute the number of matches found in the exon sequences for each matrix using minFN cut-offs. This number is defined as 100% of false positives. For every cut-off ranging from minFN to minFP, we calculate the sum of corresponding percentages for false positives and false negatives. The cut-off that gives the minimum sum is set as minSum cut-off.

INPUT AND OUTPUT

P-Match (available also as a command line executable for Linux and Windows) takes three files as input:

(i) A library of weight matrices in TRANSFAC format (see an example of one matrix entry in the Supplementary Material). Each matrix entry contains description of the nucleotide count matrix as well as an alignment of the sites that were used to build the matrix. Current web installation of P-Match uses 142 matrices of TRANSFAC public 6.0, but any other matrix.dat file of the same format can be used with the command line variant of the program.
(ii) A profile file containing a set of matrix accession numbers and two cut-offs for the *d*-score: for the whole site and for the core positions of the site.
(iii) A DNA sequence file. It can contain one or several sequences in FASTA or EMBL formats. P-Match searches sites for potential TF binding sites in these sequences and outputs all found sites in form of a table (compliant with the Match/Patch/TRANSPLORER/MatInspector output format).

The P-Match Web Interface is designed in such a way that all necessary parameters are specified by the user on the initial page. The Web Interface provides various possibilities to work with the sequence files. The content of a file can be pasted in the input window, or taken from a directory on a local computer. Files used in previous runs can be stored on the server in user-specific directory under specified names and can be reused in the next runs.

Profiles can be specified by three alternative mechanisms: (i) The user selects a taxon (vertebrate, invertebrates, plants, fungi or all) which specifies the set of corresponding weight matrices. Subsequently, the user may select a set of *d*-score cut-offs (minFN, minFP and minSUM) or may set equal cut-off values for all matrices (e.g. 0.75 for core of site and 0.85 for the whole site). (ii) The user selects one of the predefined tissue-/cell type-specific (e.g. liver-, muscle-, immune-cell) or process specific (cell cycle) profiles. By default, cut-offs in these profiles are set to minFN. (iii) The user selects one of the profiles that he/she had defined beforehand using the tool 'P-Match Profiler'.

Entry to the 'P-Match Profiler' web tool is given on the start page of the P-Match web interface. In the 'Profiler', the user can flexibly select different matrices from the whole TRANSFAC® public 6.0 matrix library (matrices which are associated with site set alignments only) and define cut-offs individually or simultaneously to all matrices in the selection and save the profile under a specified name. The user can also modify the existing profiles.

After submitting the form to the server, the P-Match program makes the search of the TF binding sites according to the given parameters. An output example of the P-Match program is shown in Figure 1. Every match found by the program is shown in a separate line of the result table. It is compliant with Match output and contains: matrix ID, position of the match, strand (+) or (−), indicating the strand orientation of the match, two *d*-scores of the match, corresponding subsequence, names of TFs associated with the matrix and the TRANSFAC site accession number of the corresponding site that provides the match. The matrix ID, factor name and the site accession number are hyperlinked to the corresponding entries of TRANSFAC database on the gene-regulation.com server. A visual representation of locations of the found matches can be generated after pressing the 'graphic' button (Figure 2b). Sites are shown above the sequence and the orientation of the '>' sign corresponds to the (+) or (−) location of the sites. The name of the matrix is given as well.

In Figure 1, we show the results of a P-Match search in the promoter of the human gene for p53 using the predefined cell cycle-specific profile. Four sites that are known in this promoter (see TRANSFAC® database) were found by P-Match (marked in the figure) along with a number of new sites.

The algorithm is implemented in C++ and the program is wrapped by a Perl script to maintain a user-friendly web interface. The P-Match tool is available at http://www.gene-regulation.com/cgi-bin/pub/programs/pmatch/bin/p-match.cgi. It is associated with the public version of TRANSFAC database (rel. pub 6.0), which is also available at this server. The gene-regulation.com server is established as a portal web site for databases and software devoted to study molecular mechanisms of gene regulation. In addition to TRANSFAC database, it contains other public releases of databases developed by BIOBASE GmbH, including TRANSPATH®, the database on signal transduction, (16); TRANSCompel®, on composite transcription regulatory elements (17); CYTOMER®, a database and ontology of human tissues, organs and cell types (18) and several other databases. It contains also a number of programs for TF site recognition such as Match, Patch as well as software program AliBaba2 (19), which applies a strategy similar to P-Match. All of these databases and software resources are very useful for further interpretation of the results obtained by P-Match program. User can attain further confirmation of the predicted sites by applying other tools to the same sequence; predicted TFs can be scanned through TRANSPATH in order to understand signal transduction pathway potentially involved in the regulation of the genes under study.

COMPARISON WITH MATCH

We performed extensive testing of P-Match on sets of genomic TF binding sites, on chromosome sequences and on artificially simulated random sequences in order to compare the recognition accuracy of this method versus classical weight matrix methods. Comparison with the weight matrix approaches such as Match™ tool was done using leave-one-out method described above. To compare the methods, we plotted the estimated values of the false negative errors versus false positive rate (estimated on the set of exon3 sequences) for the complete range of cut-off values. An example of such a comparative plot for the weight matrix V$E47_01 is given in Figure 2. In this example, the P-Match curve crosses with the Match curve at the FN value equal 0.2 showing a clear advantage over Match for the lower level of FN, which corresponds to the high sensitivity of the method, whereas for the FN level higher than 0.3, Match performs better in-site recognition accuracy. The tests show that P-Match generally provides superior recognition accuracy in the area of low false negative errors.

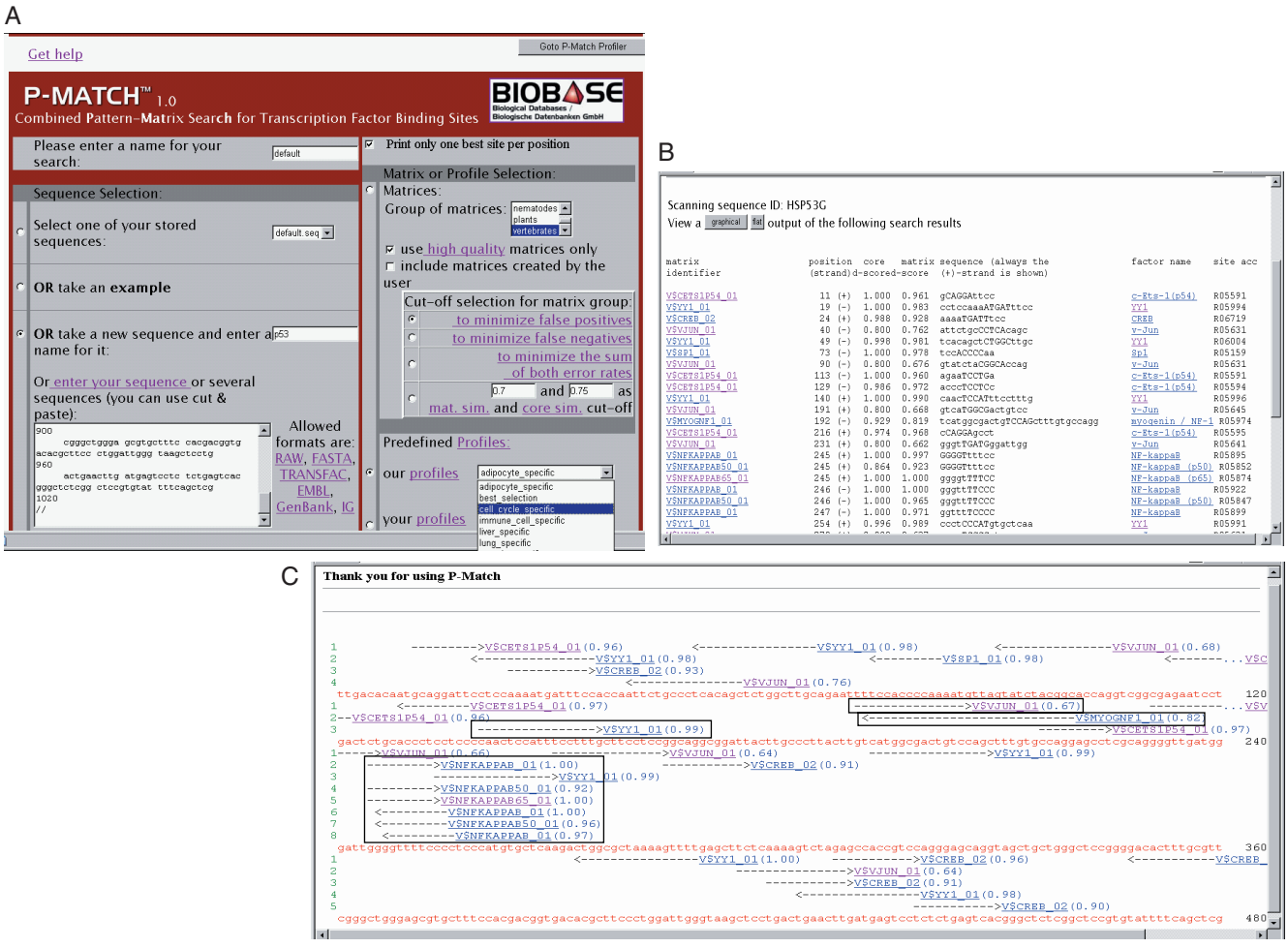

Figure 1. P-Match user interface. (**A**) Input interface. The left panel is used to paste the sequence (or several sequences) and to specify the name of the search. The right panel contains three major sections: matrix selection, cut-off selection and profile selection. (**B**) P-Match tabulated result page. Every match contains matrix ID, position of the match, strand [(+) or (−)], two *d*-scores of the match, corresponding subsequence, names of TFs associated with the matrix and the accession number of the site. Martix ID, the factor names and site accession numbers are hyperlinked with the corresponding TRANSFAC® entries. Site which are not hyperlinked are not present in the public version of TRANSFAC. (**C**) A visual representation of locations of the found matches. Sites are shown above the sequence and the orientation of the '>' sign corresponds to the (+) or (−) location of the sites. The results of P-Match search are shown for the promoter of human gene for p53 using cell cycle-specific profile. Four sites that are known in this promoter (see TRANSFAC® database) were found by P-Match (marked by a frame) along with some new sites. Here, matrix IDs are also hyperlinked with the corresponding TRANSFAC® entries.

DISCUSSION

Here, we describe a novel method for recognition of TF binding sites in gene regulatory regions of genomes. The method combines principles of pattern matching with the weighting schema of PWMs. It has clear advantages over the simple pattern matching based on Hamming distance since the new method takes into account the differential 'importance' of nucleotide positions in the site while it calculates the distance to the potential match (the match *d*-score). Such differential 'importance' of positions is a fundamental principle of the structure of TF binding sites and it is well captured by PWMs where different positions clearly diverge in their level of conservation. Often, central positions of the sites are more conserved and compose so-called site core, while the peripheral positions are less conserved. Therefore, mismatches to the pattern in the core positions of the site considered as more important than mismatches in the peripheral positions.

It is worth mentioning here that similar to Match and MatInspector (5), we use in P-Match a specific schema for calculating PWMs, which includes additional weighting of positions by information value. Multiplication of the frequencies with the information vector leads to a higher acceptance of mismatches in less conserved regions in comparison with highly conserved regions of sites. This algorithm has better performance in recognition of TF binding sites if compared with methods that do not use the information vector (8).

On the other hand, P-Match algorithm gives an advantage over the Match and other classical PWMs, which is most clearly seen in cases of highly heterogeneous site sets as it is schematically presented in Figure 3. With a classical PWM, in order to cover most of the sites in the set, we will tend to lower significantly the matrix score cut-off values, which will result in great increase of false positive rate. Whereas, implication of the individual sites as patterns with an appropriate relaxation of the *d*-score cut-offs will allow to cover the site set without too high increase of false positive rate. As we show in our comparative tests, such particularity of P-Match algorithm allows us to decrease the false positive

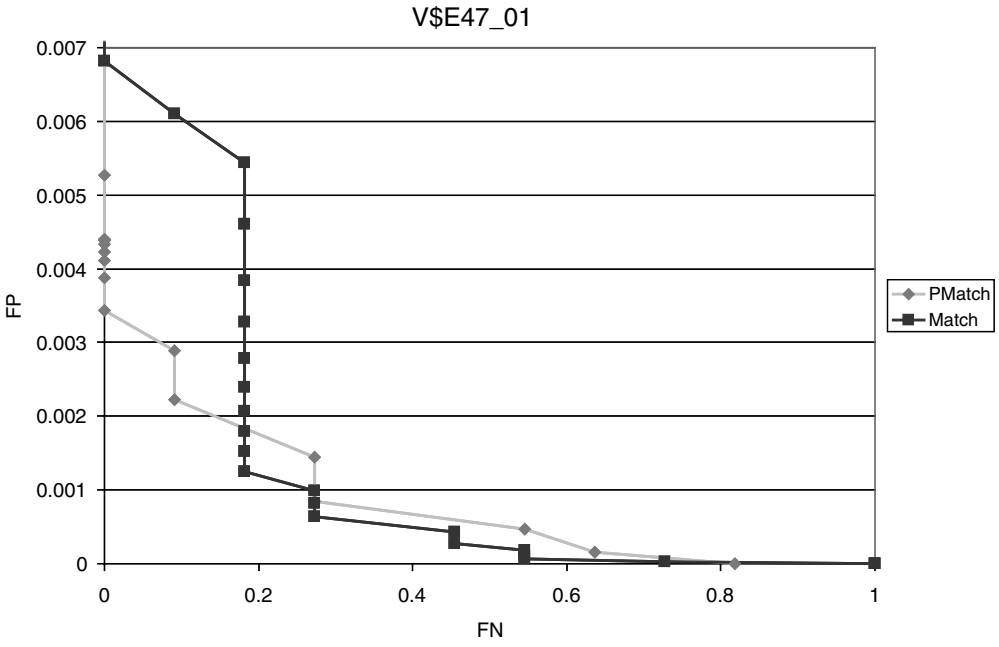

Figure 2. A comparative plot for P-Match and Match of estimated values of the false negative errors (by leave-on-out method) versus false positive (FP) rate (estimated on the set of exon3 sequences). An example of such a comparative plot for the weight matrix V$E47_01.

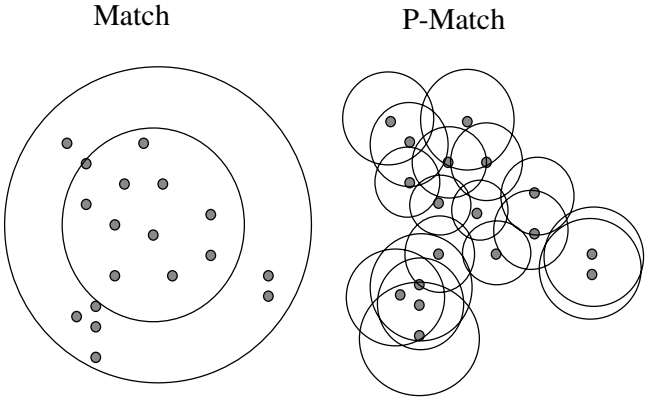

Figure 3. Schematic representation of the principles of P-Match algorithm. With the Match method, in order to cover most of the sites in the set, we will tend to lower significantly the matrix score cut-off values, which will result in great increase of false positive rate. Whereas, implication of the individual sites as patterns with an appropriate relaxation of the *d*-score cut-offs will allow to cover the site set without too high increase of false positive rate. The sizes of the circles of the Venn diagram schematically represent the number of new potential sites recognized under a certain cut-off value. In the case of P-Match, the sizes of circles become bigger while moving out of the consensus 'centre'. Such undesired behaviour of P-Match is discussed in the text.

rate of TF recognition for many matrices over two times and more in comparison to Match for the 100% sensitivity level (see table in Supplementary Material).

Nevertheless, for the 50% sensitivity level, for the most matrices, Match algorithm outperforms P-Match and gives fewer false positives. This can be explained by the tendency of P-Match algorithm to capture all the heterogeneity of the site sets including the 'weak' or, in other words, untypical representatives even under the high *d*-score cut-offs, whereas the Match algorithm focuses under high cut-offs on the 'strong' sites only and therefore produces less false positives. Theoretically, this tendency of P-Match algorithm can hamper the recognition precision, especially in cases of site sets containing many such untypical representatives, since there is much higher number of the potential sites deviating from a single consensus over the number of sites that are close to the consensus. And nucleotide weights of many of these non-consensus sites can happen to be quite close to the corresponding nucleotide weights of the untypical representatives in the set. In our future improvements of P-Match, we will consider a new scoring schema as a combination of *d*-score and the classical weight matrix score of Match algorithm.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors are indebted to Edgar Wingender (BIOBASE GmbH) for the great contribution to the preparation of the manuscript; Ingmar Reuter (BIOBASE GmbH), Evgeny Cheremushkin and Tatyana Konovalova (BioRainbow group, Novosibirsk) for help with the web interface. This work was mainly funded by BIOBASE GmbH (Wolfenbuettel, Germany). Parts of this work were supported by a grant of the European Commission (Projects: Intergenomics, COMBIO, INTAS:N.03-51-5218,TRANSISTOR), by grant BioProfile Braunschweig/Göttingen/Hannover (0 313 092). Funding to pay the Open Access publication charges for this article was provided by EU, Marie Curie Research Training Networks grant 'TRANSISTOR'.

Conflict of interest statement. None declared.

REFERENCES

1. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
2. Kel,A., Kel-Margoulis,O., Borlack,J., Tchekmenev,D. and Wingender,E. (2005) Databases and tools for *in silico* analysis of regulation of gene expression. In Borlak,J. (ed.), *Handbook of Toxicogenomics*. VCH Weinheim, pp. 253–290 (ISBN 3-527-30342-1).
3. Matys,V., Fricke,E., Geffers,R., Gößling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel Margoulis,O. *et al.* (2003) TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
4. Prestridge,D.S. (1996) SIGNAL SCAN 4.0: additional databases and sequence formats. *Comput. Appl. Biosci.*, **12**, 157–160.
5. Chen,Q.K., Hertz,G.Z. and Stormo,G.D. (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.*, **11**, 563–566.
6. Stoeckert,C.J.,Jr, Salas,F., Brunk,B. and Overton,G.C. (1999) EpoDB: a prototype database for the analysis of genes expressed during vertebrate erythropoiesis. *Nucleic Acids Res.*, **27**, 200–203.
7. Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.
8. Kel,A., Kel-Margoulis,O., Babenko,V. and Wingender,E. (1999) Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J. Mol. Biol.*, **288**, 353–376.
9. Pickert,L., Reuter,I., Klawonn,F. and Wingender,E. (1998) Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics*, **14**, 244–251.
10. Kel,A.E., Kondrakhin,Y.V., Kolpakov,Ph.A., Kel,O.V., Romashenko,A.G., Wingender,E., Milanesi,L. and Kolchanov,N.A. (1995) Computer tool FUNSITE for analysis of eukaryotic regulatory genomic sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **3**, 197–205.
11. Bucher,P. (1999) Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.*, **9**, 400–407.
12. Fickett,J.W. and Wasserman,W.W. (2000) Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.*, **11**, 19–24.
13. Conkright,M.D., Guzman,E., Flechner,L., Su,A.I., Hogenesch,J.B. and Montminy,M. (2003) Genome-wide analysis of CREB target genes reveals a core promoter requirement for cAMP responsiveness. *Mol. Cell*, **11**, 1101–1108.
14. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
15. Gößling,E., Kel-Margoulis,O.V., Kel,A.E. and Wingender,E. (2001) MATCHTM—a tool for searching transcription factor binding sites in DNA sequences. Application for the analysis of human chromosomes. In *Proceedings of the German Conference on Bioinformatics GCB'01*, October 7-10, Braunschweig, Germany, pp. 158–161.
16. Krull,M., Voss,N., Choi,C., Pistor,S., Potapov,A. and Wingender,E. (2003) TRANSPATH®: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Res.*, **31**, 97–100.
17. Kel-Margoulis,O.V., Kel,A.E., Reuter,I., Deineko,I.V. and Wingender,E. (2002) TRANSCompel®-a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.*, **30**, 332–334.
18. Michael,H., Chen,X., Fricke,E., Haubrock,M., Ricanek,R. and Wingender,E. (2004) Deriving an ontology for human gene expression sources from the CYTOMER® database on human organs and cell types. *In Silico Biol.*, **5**, 0007.
19. Grabe,N. (2000) AliBaba2: context specific identification of transcription factor binding sites. *In Silico Biol.*, **1**, 0019.